NASA TECHNICAL NOTE

NASA TN D-7932

NASA TN D-7932

COMPARISON OF SOME BIASED ESTIMATION METHODS (INCLUDING ORDINARY SUBSET REGRESSION) IN THE LINEAR MODEL

Steven M. Sidik

Lewis Research Center

Cleveland, Ohio 44135

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION • WASHINGTON, D. C. • APRIL 1975

1. Report No.
NASA TN D-7932

2. Government Accession No.

3. Recipient's Catalog No.

4. Title and Subtitle
COMPARISON OF SOME BIASED ESTIMATION METHODS (INCLUDING ORDINARY SUBSET REGRESSION) IN THE LINEAR MODEL

5. Report Date
April 1975

6. Performing Organization Code

7. Author(s)
Steven M. Sidik

8. Performing Organization Report No.
E-8180

10. Work Unit No.
506-21

9. Performing Organization Name and Address
Lewis Research Center
National Aeronautics and Space Administration
Cleveland, Ohio 44135

11. Contract or Grant No.

13. Type of Report and Period Covered
Technical Note

12. Sponsoring Agency Name and Address
National Aeronautics and Space Administration
Washington, D.C. 20546

14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

We discuss ridge, Marquardt's generalized inverse, shrunken, and principal components estimators. The discussion is with respect to the objectives of point estimation of parameters, estimation of the predictive regression function, and hypothesis testing. We find that, as the normal equations approach singularity, more consideration must be given to estimable functions of the parameters as opposed to estimation of the full parameter vector; that biased estimators all introduce constraints on the parameter space; that adoption of mean squared error as a criterion of goodness should be independent of the degree of singularity; and that ordinary least-squares subset regression is the best overall method.

17. Key Words (Suggested by Author(s))

Regression analysis
Biased estimation
Ridge regression
Linear models

18. Distribution Statement
Unclassified - unlimited
STAR Category 65 (rev.)

19. Security Classif. (of this report)
Unclassified

20. Security Classif. (of this page)
Unclassified

21. No. of Pages
46

22. Price*
$3.75

* For sale by the National Technical Information Service, Springfield, Virginia 22151

COMPARISON OF SOME BIASED ESTIMATION METHODS (INCLUDING ORDINARY SUBSET REGRESSION) IN THE LINEAR MODEL

by Steven M. Sidik

Lewis Research Center

SUMMARY

Three major types of biased estimator have recently been proposed in the literature: ridge, Marquardt's generalized inverse, and shrunken. Besides these newer biased estimators, we shall consider principal components regression and subset regression, which are also, in effect, biased estimation methods. We present the biased and unbiased estimators of the parameters in a linear model. The presentation centers on a duality of the X^TX matrix of the least-squares normal equations.

We consider biased estimators with respect to all three major objectives of a linear model analysis:

1. Estimation of parameters: (a) In a nearly singular system, the full parameter vector is essentially inestimable. However, certain linear combinations of the parameters are estimable. (b) Biased estimators all place some kind of constraint on the parameter space in order to achieve "better" estimators. (c) The decision to use mean squared error as a criterion of goodness should be made independently of the existence of multicollinearity. (d) If mean squared error is to be accepted as a criterion of goodness, only one estimator so far proposed has any proven optimality properties. (e) Because distributional information is lacking, no biased estimator provides interval estimation capability.

2. Estimation of predictive regression function: All biased estimators discussed offer the possibility of decreased mean squared error of the predictive regression function. This decrease cannot be assured (except for two special cases of principal components estimators) because it is not known how to identify the members of each class of biased estimators that provide smaller mean squared error.

3. Hypothesis testing of parameters: Only the ordinary least-squares estimators have enough of the distributional theory available to provide subset regression techniques in the original parameterization.

The overall conclusion is that ordinary least-squares estimation and subset regression methods are still the preferred methods of linear model analysis in the regression situation.

INTRODUCTION

Suppose as a result of experiment or observation you have accumulated a vector $y_{n\times 1}$ of an observed variable which you believe may be expressed as a function of a matrix $X_{n\times p}$ of p other observed predictor variables. The standard linear model assumes

$$y = Xb + e$$

where $b_{p\times 1}$ is an unknown parameter vector and where $e_{n\times 1}$ is an unobservable vector of errors of observation. Some of the objectives of an analysis of these data are

(1) Obtain an estimate of b. You may be interested in either point or interval estimates.

(2) Predict values of y at some combination of the predictor variables.

(3) Test if certain of the components of b may reasonably be set to zero.

The use of multiple linear regression analysis in achieving these goals has been intensively studied by many authors. The most admirable single summary of theory and practice is Draper and Smith (ref. 1).

Recently, some attention has been given to two aspects of regression analysis. The first aspect is the attempted improvement of point estimators where the criterion of goodness is mean squared error (Stein (ref. 2), James and Stein (ref. 3), and Sclove (ref. 4)). Sclove discusses an estimation technique which guarantees that the sum of component-wise mean squared errors of the biased estimator is smaller than that of the ordinary unbiased least-squares estimator. He also presents some further results under the restrictive condition that the terms of the equation can be ordered in importance prior to analysis. These procedures can be somewhat difficult to implement and very little is known about the distributional properties of the resulting estimators.

The second aspect of regression that has been considered recently is the problem of point estimation when there is a high degree of multicollinearity among the predictor variables (Hoerl and Kennard (refs. 5 and 6), Marquardt (ref. 7), Mayer and Willke (ref. 8), Kendall (ref. 9), and Massy (ref. 10)).

Hoerl and Kennard (ref. 5) propose a class of biased estimators called ridge estimators. Their criterion of goodness is mean squared error. The technique is relatively easy to use, and it may be shown that the class contains estimators which have smaller mean squared error than the least-squares estimator. However, they are unable to provide a well-defined and unique choice of estimator from this class. Nor have they been able to prove that their suggested procedures actually choose a member of the class which achieves smaller mean squared error. In fact, Newhouse and Oman (ref. 11) have reported some Monte-Carlo simulation results which indicate that ridge estimators do not in general perform better than least-squares estimators. There is the further drawback that the distributional properties of ridge estimators are incompletely known.

Marquardt (ref. 7) introduces a class of biased estimators called generalized inverse estimators. Mayer and Willke (ref. 8) discuss a number of classes of biased estimators called shrunken estimators. These classes have the property that they contain members with smaller mean squared error than the least-squares estimator. It is not known how to choose such members, however; and there is very little known about the distributional properties of these estimators.

Kendall (ref. 9) and Massy (ref. 10) discuss principal components regression. Principal components regression was not introduced as a method of biased estimation, but it will be shown to provide biased estimators. The method is very closely related to Marquardt's. Principal components regression was introduced for use when there is multicollinearity.

In this report we consider a method of unifying the treatment of these biased estimation methods and of unbiased least-squares estimation. The presentation centers on a duality of the X^TX matrix of the normal equations for unbiased least-squares estimation. The duality is in the sense that the spectral decomposition of X^TX into its eigenspace representation has the property of describing how well the data points are spread out in the data space. A similar decomposition of $(X^TX)^{-1}$ (or a generalized inverse of X^TX if it is singular) has the property of describing how the distribution of the estimator of b is spread out in the parameter space. We lean heavily on these decompositions to discuss the interrelationships of all these estimation methods and to describe the consequences of using them.

The report begins with a brief discussion of the objectives of a linear model analysis. The following section presents a summary of the standard estimation methods applied to linear models when X^TX is not of full rank (i.e., exactly singular) and also when X^TX is of full rank. The section also discusses the duality of X^TX in some detail. The next section summarizes the major biased estimators and some of their more important properties. In that section we also examine some of the relations among the estimators. After that we consider the mean squared error of the estimated regression function for each of the biased estimators. This is followed by a discussion of hypothesis testing procedures available for each method. The last section presents two numerical examples to illustrate the results developed in this report.

OBJECTIVES OF LINEAR MODEL ANALYSES

The model that we are dealing with is

$$y = Xb + e \tag{1}$$

where

y an $n \times 1$ vector of observations

b a $p \times 1$ vector of unknown parameters

X an $n \times p$ matrix of known values of p predictor variables for each of n observations

e an $n \times 1$ vector of random errors which we assume has $N(0, \sigma^2 I)$ distribution

In terms of this model we generally wish to consider any one or more of the following:

(1) Find an estimate of b. The components of b might represent either physical constants or perhaps just empirical rates of change. We are usually concerned only with point estimates but occasionally desire interval estimates.

(2) Predict a value of y for some point $X_{0(1\times p)}$ of the space of the predictor variables. This is often the most important objective.

(3) Test if certain of the components of b may reasonably be set to zero (or some other specified constant). If the parameters represent some physical constants, such tests may provide evidence for or against some theory. For purposes of prediction or control, if certain components of b can be set to zero, this implies the corresponding predictor variables have no effect on y and hence may be ignored. This provides simplicity and often economy. It also often reduces the variance of the estimated predicting equation.

Most of the previous authors have considered only one of these objectives when developing biased estimators. We will consider all three.

LINEAR MODEL ESTIMATION PROBLEM

Our model is that described by equation (1). For this model, it is well known that either least-squares or maximum-likelihood arguments lead to the minimum-variance unbiased estimator for b as any solution b^0 to the normal equations

$$X^T X b^0 = X^T y \tag{2}$$

In fact, $X^T X$ will be either singular or nonsingular; but in practice we may consider $X^T X$ to be singular, nearly singular, or nonsingular. If $X^T X$ is singular, there are many solutions b^0 to equation (2). The section Models Not of Full Rank describes estimation concepts and procedures for this situation. If $X^T X$ is nonsingular, there is exactly one solution to equation (2). If $X^T X$ is nearly singular, then we might expect that there will be difficulty in deciding whether we have a solution or many poorly defined solutions. A nearly singular X^TX is a symptom of multicollinearity. The section Models of (Just Barely) Full Rank describes estimation concepts and procedures for the X^TX nonsingular and nearly singular situations. The last section describes the spectral decomposition of X^TX and its interpretation.

Models Not of Full Rank

The study of linear models based upon generalized inverses for the X^TX singular situation has been most lucidly presented by Searle (refs. 12 and 13). We use his notation and an example given by Federer (ref. 14), which is discussed in chapter 5 of Searle (ref. 13), to review the basics.

As stated previously, when X^TX is singular, there is no unique solution b^0 to equation (2). We begin by letting $G = (X^TX)^+$ be a generalized inverse of X^TX. That is, G satisfies

$$G = (X^TX)^+ \tag{3}$$

and

$$(X^TX)G(X^TX) = (X^TX) \tag{4}$$

Let

$$H = (X^TX)^+(X^TX) \tag{5}$$

Then

$$b^0 = (X^TX)^+X^Ty \tag{6}$$

is a solution to equation (2) (not unique) such that the expectation of b^0 is

$$E(b^0) = Hb \tag{7}$$

and the variance of b^0 is

$$V(b^0) = G(X^TX)G^T\sigma^2 \tag{8}$$

Since G and hence H are not unique, there are infinitely many solutions b^0 to equation (2). An estimable function of the parameter vector b is any linear function of b for which an estimator can be found from b^0 which is invariant to whatever the choice of G.

Searle (ref. 12) has shown that all estimable functions are of the form

$$w^T Hb$$

for arbitrary choice of w. There are, at most, r linearly independent estimable functions where $r = \text{rank}(X^T X)$. The estimator for $w^T Hb$ is given by $w^T GX^T y$, and this is unbiased for $w^T Hb$. The variance is

$$V(w^T G^T X^T y) = w^T GX^T XG^T w\sigma^2$$

The example we consider is discussed in Searle (ref. 13) where weights are presented for six rubber plants, three of which are normal, two of which are off-type, and one of which is an aberrant. The data are presented in table I. The model we consider is

$$y_{ij} = \mu + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + e$$

where

$$X_i = \begin{cases} 1 & \text{if the plant is of the } i^{th} \text{ type} \\ 0 & \text{otherwise} \end{cases}$$

We thus have the model represented as in equation (1), where

$$y = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{31} \end{bmatrix} = \begin{bmatrix} 101 \\ 105 \\ 94 \\ 84 \\ 88 \\ 32 \end{bmatrix} \qquad b = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \tag{9}$$

In this example we have

$$(X^TX) = \begin{bmatrix} 6 & 3 & 2 & 1 \\ 3 & 3 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

and it is seen that X^TX is singular and of rank 3. A generalized inverse of X^TX is

$$G = (X^TX)^+ = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1/3 & 0 & 0 \\ 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{10}$$

and for this choice of G we have

$$H = G(X^TX) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \tag{11}$$

Hence, all estimable functions are of the form

$$w^THb = (w_1 + w_2 + w_3)\mu + w_1\alpha_1 + w_2\alpha_2 + w_3\alpha_3 \tag{12}$$

and there are, at most, three linearly independent choices of w. The unbiased estimates of w^THb are given by $w^TGX^Ty = w_1\overline{y}_{1.} + w_2\overline{y}_{2.} + w_3\overline{y}_{3.}$. Three reasonable choices for independent estimable functions are provided in table II.

To emphasize, the major point under discussion is that there is no unique solution to the normal equations. We could make it unique by imposition of a constraint such as $\alpha_1 + \alpha_2 + \alpha_3 = 0$. In place of such constraints, it is often more useful to concentrate on choosing the w_i values that lead to meaningful estimable functions.

Models of (Just Barely) Full Rank

When X^TX is nonsingular, the estimation of b and the description of the distribution of its estimator are more straightforward. For in this case it is well known that

$$\hat{b} = (X^TX)^{-1}X^Ty$$

is unique and is the minimum-variance unbiased estimator. Also $\hat{b}$ has the following distribution:

$$\hat{b} \sim N\left[b,\ \sigma^2 (X^T X)^{-1}\right]$$

From properties of the multivariate normal distribution, it is well known that a set of linear combinations of $\hat{b}$, say $K^T\hat{b}$, has the following distribution:

$$K^T\hat{b} \sim N\left[K^T b,\ \sigma^2 K^T (X^T X)^{-1} K\right] \tag{13}$$

Now, to consider what problems arise as we slowly bridge the gap from $X^T X$ nonsingular to singular, suppose we modify the previous example. We just barely remove it from the singular setting and perform a regression analysis. Suppose we perform an experiment to study the abrasion resistance of rubber as a function of the amount of three particular additives. Let X_i denote the amount in pounds of the i^{th} additive which is loaded with an approximately 1000-pound charge to the chemical reactor which produces the rubber.

The proposed model is

$$y = Xb + e$$

where

$$X = \begin{bmatrix} 1 & 0.99 & 0 & 0 \\ 1 & 1.00 & 0 & 0 \\ 1 & 1.01 & 0 & 0 \\ 1 & 0 & 0.99 & 0 \\ 1 & 0 & 1.01 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

$$y^T = \left[y_1,\ y_2,\ y_3,\ y_4,\ y_5,\ y_6\right]$$

$$= [101,\ 105,\ 94,\ 84,\ 88,\ 32]$$

$$b^T = \left[\mu,\ \alpha_1,\ \alpha_2,\ \alpha_3\right]$$

and e is defined as previously (following eq. (1)). In this example we have

$$X^TX = \begin{bmatrix} 6 & 3 & 2 & 1 \\ 3 & 3.0002 & 0 & 0 \\ 2 & 0 & 2.0002 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

and it is evident that X^TX is not singular. Yet, it is nearly so. Let λ_i denote the eigenvalues of X^TX. When these data were submitted to NEWRAP (ref. 15), the following eigenvalues were calculated:

$$\lambda_1 = 8.41888$$

$$\lambda_2 = 2.38695$$

$$\lambda_3 = 1.19444$$

$$\lambda_4 = 0.00010$$

Since $\lambda_4 \approx 0.0$, we see X^TX is nearly singular.

The following parameter estimates were obtained

$$\left.\begin{aligned} \hat{\mu} &= 168.002 \\ \hat{\alpha}_1 &= -68.0212 \\ \hat{\alpha}_2 &= -81.9742 \\ \hat{\alpha}_3 &= -136.002 \end{aligned}\right\} \qquad (14)$$

with a covariance matrix of $(X^TX)^{-1}\sigma^2$ where

$$(X^TX)^{-1} = \begin{bmatrix} 2500.38 & -2500.21 & -2500.13 & -2500.38 \\ -2500.21 & 2500.38 & 2499.96 & 2500.21 \\ -2500.13 & 2499.96 & 2500.38 & 2500.13 \\ -2500.38 & 2500.21 & 2500.13 & 2501.38 \end{bmatrix} \qquad (15)$$

It is evident that X^TX is formally of rank 4 although it is essentially of rank 3 and that the resulting $(X^TX)^{-1}$ matrix indicates a large variance in the parameter estimates.

However, let us recall the estimable functions discussed in the previous section. Namely, let us compute

$$\left.\begin{aligned} \hat{\alpha}_1 - \hat{\alpha}_2 &= 13.9530\ (14) \\ \hat{\alpha}_2 - \hat{\alpha}_3 &= 54.0278\ (54) \\ \hat{\mu} + \frac{1}{3}(\hat{\alpha}_1 + \hat{\alpha}_2 + \hat{\alpha}_3) &= 72.6695\ \left(72\,\frac{2}{3}\right) \end{aligned}\right\} \tag{16}$$

The numbers in parentheses are the corresponding values from the analysis of variance (ANOVA) situation. Allowing for the fact that X was slightly changed to provide nonsingularity, the agreement is admirable.

Although the variances of the raw estimates $\hat{\mu}$, $\hat{\alpha}_1$, $\hat{\alpha}_2$, and $\hat{\alpha}_3$ are quite large, let us consider the variances of the linear combinations of parameters in equation (16). The linear combinations are defined by $K^T\hat{b}$, where

$$K = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & \frac{1}{3} \\ -1 & 1 & \frac{1}{3} \\ 0 & -1 & \frac{1}{3} \end{bmatrix}$$

From equation (13) the covariance matrix of $K^T\hat{b}$ is given by $K^T(X^TX)^{-1}K\sigma^2$. But

$$K^T(X^TX)^{-1}K = \begin{bmatrix} 0.82 & -0.33 & 0.05 \\ -0.33 & 1.50 & 0.17 \\ 0.05 & -0.17 & 0.53 \end{bmatrix}$$

It is thus evident that, even though the full parameter vector is quite ill determined, the linear combinations of the parameters corresponding to the estimable functions of the ANOVA example are well determined. This is, of course, not at all surprising.

Duality of X^TX

The normal equations matrix X^TX plays the central role in linear model estimation and hypothesis testing. We first note that X^TX has a spectral decomposition (ref. 16, p. 36) or representation as

$$X^TX = \sum_{i=1}^{p} \lambda_i S_i S_i^T \tag{17}$$

where $\lambda_1 \geq . . . \geq \lambda_p \geq 0$ are the eigenvalues of X^TX and S_i are the corresponding normalized eigenvectors of X^TX. If r is the rank of X^TX, then we have a similar decomposition (ref. 16, p. 184) of $(X^TX)^+$ as

$$(X^TX)_r^+ = \sum_{i=1}^{r} \frac{1}{\lambda_i} S_i S_i^T \tag{18}$$

If $r = p$, then

$$(X^TX)_p^+ = (X^TX)^{-1} = \sum_{i=1}^{p} \frac{1}{\lambda_i} S_i S_i^T \tag{19}$$

An extremely important point to note is that the spectral representations of equations (17) and (18) are not invariant under linear transformations of X. Invariance may be attained by assuming that we always consider the linear model in its correlation form. That is, we will in the remainder of this report assume that X^TX has all its diagonal elements equal to unity. This may always be achieved by a simple linear transformation corresponding to changes in scale and/or location of the original predictor variables. Since the correlation matrix is always invariant with respect to changes of location and scale, the spectral decomposition of the correlation matrix will be unique and well defined. Thus, as long as we use the convention of reducing the model to correlation form, the representation is invariant to changes in location and scale of the predictor variables.

The spectral decomposition of X^TX by equation (17) indicates how and how well the variables' space is spanned by the experiment. Namely, if $\lambda_i = 1.0$ for all i, then in a sense the variables' space is perfectly spanned. If $\lambda_1 >> \lambda_p$, then the variables' space is not well spanned. In fact, X_0S_1 represents the linear subspace (or linear combination) of predictor variables which is spanned the best. And X_0S_p represents the linear combination of variables which is most poorly spanned. In fact, if $\lambda_p = 0$, then X_0S_p is not spanned at all. These considerations are discussed by Kendall and Stuart (ref. 17, p. 287).

In order to illustrate the preceding, consider the following two-dimensional example.

Suppose the data points observed are as plotted in figure 1. Assume that the dashed line ξ_1 is the line $x_1 - x_2 = 0$ in the variables' space and that ξ_2 is the line $x_1 + x_2 = 0$. Assume also that the two extreme points along ξ_2 are equally distant from $x_1 - x_2 = 0$.

It is immediately seen that the observations are much more spread out along ξ_1 (i.e., $x_1 - x_2 = 0$) than along ξ_2 (i.e., $x_1 + x_2 = 0$). For such a situation we would find that $\lambda_1 > \lambda_2$; and we would say that X_0S_1 is well spanned, while X_0S_2 is poorly spanned.

Turning to consideration of the parameter space, it is well known that the least-squares estimator $\hat{b}$ (under normal distribution theory) follows the normal distribution $N\left[(X^TX)_r^+(X^TX)b,\ \sigma^2(X^TX)_r^+\right]$. From reference 13 (p. 185) we have that, for a linear combination of the estimates $w^T\hat{b}$,

$$V(w^T\hat{b}) = w^T(X^TX)_r^+w\sigma^2$$

It can be shown (ref. 16, p. 501) that the choice of w which minimizes the variance of $w^T\hat{b}$ is $w = S_1$ and that this variance is

$$V\left(S_1^T\hat{b}\right) = S_1^T(X^TX)_r^+S_1\sigma^2 = \frac{\sigma^2}{\lambda_1}$$

The choice of w which maximizes the variance of $w^T\hat{b}$ is $w = S_p$ and

$$V\left(S_p^T\hat{b}\right) = \frac{\sigma^2}{\lambda_p} \qquad \text{(assuming } p = r\text{)}$$

Thus, $S_1^T\hat{b}$ describes the most determined linear combination of the parameters, while $S_p^T\hat{b}$ describes the least determined. In fact, if $\lambda_p = 0$, then $S_p^T\hat{b}$ is nonestimable and hence not determined at all. An interpretation of this is that $S_p^T\hat{b}$ has infinite variance.

SOME BIASED ESTIMATORS

We now describe briefly some of the biased estimators that have been proposed and some of their most important properties.

Ridge Estimators

There are two forms of ridge estimator proposed by Hoerl and Kennard (ref. 5). One is a general form and the other is a form more useful for applications.

In the general form, we begin with the model of equation (1)

$$y = Xb + e$$

We may represent X^TX as $X^TX = P\Lambda P^T$, where P is the orthogonal matrix whose columns are the normalized eigenvectors of X^TX and Λ is the diagonal matrix of eigenvalues. Considering the transformation to new predictor variables defined by

$$W = XP$$

and the model

$$y = Wa + e$$

we have

$$a = P^Tb$$

$$W^TW = \Lambda$$

$$a^Ta = b^Tb$$

The general ridge estimation procedure is defined by the family of estimators indexed by the parameters $k_i \geq 0$

$$a* = (W^TW + K)^{-1}W^Ty \tag{20}$$

where the matrix K is defined by

$$K = (\delta_{ij}k_i)$$

When all $k_i = 0$, $a*$ is the ordinary least-squares estimator and is unbiased. When any $k_i > 0$, the resulting estimator for a is biased. We define the mean squared error of $a*$ as

$$M(K) = E\left[(a* - a)^T(a* - a)\right]$$

It may be shown (see Hoerl and Kennard, ref. 5) that the choice of $k_i = \sigma^2/a_i^2$ will minimize $M(K)$ among the class of estimators defined by equation (20). Unfortunately, in order to utilize this optimal choice of k_i, one must know both σ^2 and a_i^2. But if this information is available, there is no estimation problem. It may also be noted that in this canonical representation, the matrix $(W^TW + K)$ is diagonal. Thus, the estimation of a reduces to the independent estimation of the components of a.

In the preceding form of ridge regression there are p k's to choose. In order to provide a reasonably tractable method of analysis, Hoerl and Kennard consider also the model

$$y = Xb + e$$

where it is assumed only that the X matrix is scaled such that X^TX has diagonal elements equal to unity. That is, they consider the model in its correlation form. The more specific form of ridge regression is then defined by the family of estimators

$$\hat{b}(k) = (X^TX + kI)^{-1}X^Ty \qquad k \geq 0 \tag{21}$$

Hoerl and Kennard have shown that this family has the property that there always exists a $k > 0$ such that

$$M(k) = E\left\{[\hat{b}(k) - b]^T[\hat{b}(k) - b]\right\}$$

$$= \sigma^2 \sum_i \frac{\lambda_i}{(\lambda_i + k)^2} + k^2b^T(X^TX + kI)^{-2}b$$

is minimized. It may be shown also that $M'(0) < 0$. There is no way yet developed for determining the "best" k. Hoerl and Kennard's suggested procedure for choosing an estimator from this class is to plot the elements of $\hat{b}(k)$ for a number of values of k between 0 and 1. The "best" value of k is then subjectively chosen as the point at which these curves begin to stabilize. They have not shown that this procedure even guarantees a reduction in $M(k)$ let alone a minimum.

In what follows, we will be interested in the quantities $S_i^T\hat{b}(k)$ as indicated in the section Duality of X^TX. We obtain immediately

$$
\begin{aligned}
S_i^T\hat{b}(k) &= S_i^T(X^TX + kI)^{-1}X^Ty \\
&= S_i^T\left(\sum_j \frac{1}{\lambda_j + k} S_jS_j^T\right)X^Ty \\
&= \frac{1}{\lambda_i + k} S_i^TX^Ty \\
&= \frac{v_i}{\lambda_i + k} \qquad (22)
\end{aligned}
$$

where $v_i = S_i^TX^Ty$. We also have

$$
\begin{aligned}
E\left[S_i^T\hat{b}(k)\right] &= \frac{1}{\lambda_i + k} S_i^TX^TE(y) \\
&= \frac{\lambda_i}{\lambda_i + k} S_i^Tb \qquad (23)
\end{aligned}
$$

and

$$
\begin{aligned}
V\left[S_i^T\hat{b}(k)\right] &= \frac{1}{(\lambda_i + k)^2} S_i^TX^T(\sigma^2I)XS_i \\
&= \frac{\lambda_i\sigma^2}{(\lambda_i + k)^2} \qquad (24)
\end{aligned}
$$

Thus, for any nonzero k, $S_1^T\hat{b}(k)$ is the least biased linear combination of the estimator and $S_p^T\hat{b}(k)$ is the most biased. Also $S_1^T\hat{b}(k)$ has the least reduced variance, and $S_p^T\hat{b}(k)$ has the most reduced variance. Thus, the best determined linear combinations of the parameter estimates are the least modified, while the least determined are the most modified. In effect, as k increases, those $S_i^T\hat{b}(k)$ corresponding to small λ_i are rapidly driven to zero. The estimators are thus constrained estimators.

Marquardt's Generalized Inverse Estimators

Marquardt (ref. 7) discusses a method of applying generalized inverses to biased estimation. He also considers some relations among these estimators, ridge estimators, and nonlinear estimation. He considers the model of equation (1) where the X matrix has been scaled so that X^TX is in the correlation form. His family of estimators is indexed by a parameter ρ where $0 \leq \rho \leq p$. The family is defined by

$$\hat{b}(\rho) = (X^TX)^+_\rho X^T y \tag{25}$$

The matrix $(X^TX)^+_\rho$ is defined as follows: Let $\rho^* = [\rho]$ denote the greatest integer in ρ and $d\rho = \rho - \rho^*$. Then $(X^TX)^+_\rho$ is defined as

$$\begin{aligned}(X^TX)^+_\rho &= \sum_{j=1}^{\rho^*} \frac{1}{\lambda_j} S_j S_j^T + \frac{d\rho}{\lambda_{\rho^*+1}} S_{\rho^*+1} S^T_{\rho^*+1} \\ &= G_\rho \end{aligned} \tag{26}$$

As the notation is meant to indicate, $(X^TX)^+_\rho$ is closely related to a generalized inverse of X^TX. In fact, if $r = \text{rank}(X^TX)$, then $(X^TX)^+_r$ is the Moore-Penrose generalized inverse of X^TX and is unique. An important point to note is that it is well known that the Moore-Penrose generalized inverse yields the minimum-norm solution to the normal equations (ref. 18, p. 50).

Marquardt's estimators thus provide a sort of minimum-norm solution to the normal equations. He pursues this idea in some depth in reference 7. He also shows that there always exists a $0 < \rho < p$ such that

$$M(\rho) = E\left\{[\hat{b}(\rho) - b]^T[\hat{b}(\rho) - b]\right\}$$

is minimized. It is also shown that $M'(p) > 0$ so that the mean squared error of $\hat{b}(\rho)$ is initially decreasing as ρ decreases from p. As with the ridge estimators, there is no way yet developed for determining the "best" ρ.

With X scaled so that X^TX is in the correlation form, Marquardt labels the diagonal elements of $(X^TX)^+_\rho$ as variance inflation factors. His suggested analytical procedure is to consider several estimates $\hat{b}(\rho)$ for ρ between p and 0. He suggests the rule of thumb that an acceptable value of ρ is one such that the maximum variance inflation factor should usually be larger than 1.0 but certainly not as large as 10.0.

Marquardt has not been able to show that this procedure even guarantees a reduction in $M(\rho)$ let alone a minimum.

For these estimators we have

$$S_i^T\hat{b}(\rho) = S_i^T(X^TX)_\rho^+X^Ty = \begin{cases} 0 & \rho \leq i-1 \\ \dfrac{d\rho}{\lambda_i} S_i^TX^Ty & i-1 < \rho \leq i \\ \dfrac{1}{\lambda_i} S_i^TX^Ty & i < \rho \end{cases} \tag{27}$$

It may rather easily be shown that

$$E\left[S_i^T\hat{b}(\rho)\right] = \begin{cases} 0 & \rho \leq i-1 \\ (d\rho)S_i^Tb & i-1 < \rho \leq i \\ S_i^Tb & i < \rho \end{cases} \tag{28}$$

and

$$V\left[S_i^T\hat{b}(\rho)\right] = \begin{cases} 0 & \rho \leq i-1 \\ \dfrac{(d\rho)^2\sigma^2}{\lambda_i} & i-1 < \rho \leq i \\ \dfrac{\sigma^2}{\lambda_i} & i < \rho \end{cases} \tag{29}$$

From equations (28) and (29) we see the following behavior as ρ decreases from $\rho = p$: The $S_i^T\hat{b}(\rho)$ are successively set to zero in order of increasing λ_i. The best determined linear combinations of the parameter estimates are the last to be set to zero, while the least determined are the first to be set to zero.

Shrunken Estimators

Mayer and Willke (ref. 8) discuss several families of biased estimators which may all be labeled shrunken estimators. They consider the model of equation (1), $y = Xb + e$, but do not require that X^TX be in correlation form. Each family of estimators is indexed by a parameter $1 \geq c \geq 0$ and defined by

$$\hat{b}(c) = c(X^TX)^{-1}X^Ty = c\hat{b} \tag{30}$$

where $\hat{b}$ is the ordinary unbiased least-squares estimator. If the constant c is a scalar fixed in advance of the analysis, then $\hat{b}(c)$ is called a deterministically shrunken estimator. If $c = f(\hat{b}^T\hat{b})$ is a scalar function of the least-squares estimator, then $\hat{b}(c)$ is called a stochastically shrunken estimator.

It may be shown that there does exist a value of $0 < c < 1$ such that

$$M(c) = E\left\{[\hat{b}(c) - b]^T[\hat{b}(c) - b]\right\}$$

is minimized. Consider the stochastically shrunken estimator $\hat{b}(c)$, where

$$\left.\begin{aligned} c &= \left[1 - \xi S^2(\hat{b}^T\hat{b})^{-1}\right] \\ S^2 &= y^Ty - \hat{b}^T(X^TX)^{-1}\hat{b} \\ p &\geq 3 \\ \text{and} \qquad 0 < \xi &< 2(p - 2)(n - p + 2)^{-1} \end{aligned}\right\} \tag{31}$$

This estimator is one discussed by Sclove (ref. 4). Then if we define

$$W[\hat{b}(c)] = E\left\{[\hat{b}(c) - b]^T[\hat{b}(c) - b]\right\}$$

it was shown by Sclove (based on results of refs. 2 and 3) that

$$\xi = \xi_0 = \frac{p - 2}{n - p + 2}$$

minimizes $W[\hat{b}(c)]$. This is the only biased estimator known to this author for which a choice of biased estimator can be explicitly given that guarantees a reduction in mean squared error.

Mayer and Willke develop another family of stochastically shrunken estimators indexed by $\delta \geq 0$

$$\hat{b}(\delta) = \delta \hat{b}\hat{b}^T (I + \delta \hat{b}\hat{b}^T)^{-1} \hat{b} .$$

Few properties of this estimator are known. Mayer and Willke suggest the user plot the components of $\hat{b}(\delta)$ as a function of δ and choose that δ where the curves begin to stabilize. It is not known whether this procedure provides an estimator with smaller mean squared error than $\hat{b}$.

For purposes of comparison to the other biased estimation procedures we will consider only the deterministically shrunken estimator $\hat{b}(c) = c\hat{b}$. For this choice we have

$$S_i^T \hat{b}(c) = cS_i^T \hat{b} = \frac{c}{\lambda_i} S_i^T X^T y \tag{32}$$

$$E\left[S_i^T \hat{b}(c)\right] = cS_i^T b \tag{33}$$

and

$$V\left[S_i^T \hat{b}(c)\right] = \frac{c^2 \sigma^2}{\lambda_i} \tag{34}$$

From equations (33) and (34) we observe that, unlike the ridge and generalized inverse estimators, all linear combinations of the parameter estimates are driven toward zero proportionately and that the variances are also proportionately reduced.

Principal Components Estimators

Principal components regression (or canonical regression as it is sometimes called) does not appear to be widely used in the physical sciences although it is apparently used in economics and the social sciences. The description of principal components regression will be primarily along the lines of Kendall (ref. 9) and Massy (ref. 10) although there may be some differences.

We begin with the model of equation (1), $y = Xb + e$, and make an orthogonal transformation of this model to that of

$$y = Wa + e$$

where

$$X^TX = P\Lambda P^T$$

$$P^TP = PP^T = I$$

$$W = XP$$

and

$$a = P^Tb$$

The vector of parameters a is then estimated by

$$\begin{aligned} \hat{a} &= (W^TW)^{-1}W^Ty \\ &= \Lambda^{-1}W^Ty \end{aligned} \tag{35}$$

Since Λ is the diagonal matrix of eigenvalues of X^TX, each component of $\hat{a}$ is independent. This is the principal reason for the transformation. Massy discusses two methods of obtaining estimators for b from this form of the model. The first method consists of setting to zero the components of $\hat{a}$ which correspond to "small" eigenvalues. Let $\hat{a}(d)$ denote the resulting estimate of a. (That is, the components of $\hat{a}(d)$ are equal to zero for those components with small λ_i and are equal to the corresponding components of $\hat{a}$ for those with large λ_i.) We then define the estimator for b as

$$\hat{b}(d) = P\hat{a}(d)$$

It should be noted that this is precisely Marquardt's generalized inverse estimator when ρ is an integer. Kendall also uses this method. We will not refer to this method as principal components estimation.

The second method discussed by Massy is to test the components of $\hat{a}$ for significance from zero. This may be done, for instance, by the usual t-tests. Other methods are discussed by Kennedy and Bancroft (ref. 19) and Holms (ref. 20). In this case we will obtain an estimator for b which we will still denote as $\hat{b}(d)$ and define as

$$\hat{b}(d) = P\hat{a}(d) \tag{36}$$

For either of these methods we obtain

$$S_i^T\hat{b}(d) = S_i^T P\hat{a}(d) = \begin{cases} \frac{1}{\lambda_i} S_i^T X^T y & \hat{a}_i(d) \neq 0 \\ 0 \quad \hat{a}_i(d) = 0 \end{cases} \tag{37}$$

depending upon whether the i^{th} component of $\hat{a}(d)$ is nonzero or zero. The expectations and variances of these quantities are dependent upon the method for choosing $\hat{a}(d)$. In particular, they will be the same as for Marquardt's estimators if Massy's first method is used. If his second method is used, these quantities would be difficult to obtain.

It should be noted that obtaining $\hat{a}(d)$ by the second method will usually be on good statistical footing because the independence of the components of $\hat{a}$ permits easier analysis than the nonindependent situations more commonly found in regression.

Discussion

There are five points we shall touch upon in this section:

(1) In point estimation, consideration should be given to the estimable functions of the parameters.

(2) Biased point estimators all place some form of constraint upon the parameter space.

(3) The decision to use mean squared error as a criterion for the goodness of estimators should be based on the objectives of the data analysis and be made independent of the condition of X^TX.

(4) If mean squared error is to be the criterion of goodness, then Sclove's estimator is the only one which has any proven optimality properties.

(5) Because of the lack of distributional information, no biased estimator provides interval estimation capability.

We now consider these points in more detail.

Estimability. - As the section LINEAR MODEL ESTIMATION PROBLEM pointed out, when the X^TX matrix is precisely singular there is no unique estimator for b. Instead we should take the approach of examining estimable functions of the parameters. The major problem in application would be that of choosing meaningful estimable functions. There seems to be no good reason for abandoning this approach as soon as X^TX becomes the least bit nonsingular. What seems most reasonable is that less and less attention need be paid to estimability considerations as X^TX deviates more and more from singularity. The transition of viewpoint should be smooth rather than a quantum leap.

Constraints. - Each method of biased estimation (including subset regression - which is also, of course, a form of biased estimation) introduces constraints on the parameter space. In the analysis of variance, the model is typically over parameterized. But it is over parameterized in such a manner that certain constraints are "natural." In the regression situation, "natural" constraints seem unlikely to present themselves.

The constraints imposed by biased estimators are as follows: The generalized inverse estimators of Marquardt and the principal components estimators of both methods drop linear subspaces out of the parameter space. The subspaces are defined by the constraints that $S_i^T\hat{b}(\rho) = 0$ or $S_i^T\hat{b}(d) = 0$ (eqs. (27) and (37)). Which particular subspaces are dropped out, of course, depends upon how it is decided to drop them. We then choose minimum-norm solutions in the subspaces remaining. Ridge estimators have a very closely related behavior. From the expressions for $S_i^T\hat{b}(k)$ (eq. (22)), it may be seen that for the linear combinations corresponding to large λ_i the addition of k to the denominator has a lesser effect than for those with small λ_i. Each combination, however, is driven to zero, with those with small λ_i getting there quicker. The shrunken estimators are more difficult to characterize, but there are constraints nonetheless.

The final comment along these lines might be that ordinary subset regression techniques also provide biased estimators and also impose constraints. The constraints are that certain components of $\hat{b}$ are set to zero and as such have an extremely clear interpretation.

Condition of X^TX. - Following the comments about estimability and constraints, it seems that biased estimation is not necessarily a good means of removing the symptoms of multicollinearity. It has been belabored in references 5 to 8 that one consequence of multicollinearity is that the estimators tend to be "too large." Since the least-squares estimator is unbiased, it should also possess to some degree the tendency for estimators to be "too small." In fact, much of the confusion can probably be attributed to misinterpretations of the type found on page 58 of Hoerl and Kennard (ref. 5). They state, "However, the relationships in section 2 show that on the average, the distance from $(\hat{b})$ to (b) will tend to be large if there is a small eigenvalue of X^TX. In particular, the worse the conditioning of X^TX, the more $\hat{b}$ can be expected to be too long." The first statement is not correct since unbiasedness implies the average distance is zero.

Thus, mean squared error has been proposed as a criterion of goodness when X^TX is ill conditioned. But it is either a good criterion or a bad criterion independent of X^TX.

Sclove's estimator. - Of all the biased estimators, only Sclove's can claim a guaranteed improvement over the least-squares estimator. Principal components estimation based on significance testing should offer possibilities for improvement. This is only conjecture at this point. Ridge estimation, generalized inverse estimation, and shrunken estimation do not show how to choose better estimators. Their proponents show only that better ones exist. For Bayesian statisticians, there are comments in references 5 and 7 which indicate these might be useful estimators. This is so because a Bayesian statistician assumes the user is able to specify prior information about σ^2 and b.

Interval estimation. - Until such time as the distributional properties of the biased estimators are discovered, they cannot be used to provide interval estimators. Only least-squares estimation without subset regression provides the necessary distribution theory.

MEAN SQUARED ERROR OF ESTIMATED REGRESSION FUNCTION

One of the primary objectives of a linear model analysis is to provide a predictive equation. Since the biased estimators discussed in the previous section all show that there exist members in their respective classes with smaller mean squared error, it seems likely that their use could also provide predictive equations with smaller mean squared error. It will be seen that, in fact, this can be shown to be true for most of the biased estimators. In order to provide a common reference point, we also present the mean squared error of the least-squares predictive equation. Since this latter predictor is unbiased, the mean squared error reduces to the variance.

Least Squares

For any estimator $\hat{b}^*$ of b, we will consider the use of $\hat{b}^*$ to predict the estimated regression function at a set of values of the predictor variables denoted by X_0 (assumed to be a $1 \times p$ vector). We recall the model

$$y = Xb + e$$

For an estimator $\hat{b}^*$ based on this model and data, the predicted regression function at X_0 is

$$\hat{y}_0 = X_0\hat{b}*$$

and the mean squared error of this predicted regression is denoted by

$$M_p\left(\hat{y}_0 \middle| \hat{b}*\right) = E\left[(\hat{y}_0 - X_0 b)^T(\hat{y}_0 - X_0 b)\right]$$

$$= E\left[e^T X(X^T X)^{-1} X^T X_0^T X_0 X(X^T X)^{-1} X^T e\right] + 0$$

$$= \gamma_1(\hat{b}*) + \gamma_2(\hat{b}*) \tag{38}$$

where $\gamma_1(\hat{b}*)$ corresponds to the variance and $\gamma_2(\hat{b}*)$ to the bias squared. For the least-squares estimator, $\hat{b}* = \hat{b}$ is unbiased and consequently $\hat{y}_0 = X_0\hat{b}$ is unbiased for $X_0 b$. Thus,

$$\hat{b} = (X^T X)^{-1} X^T y = b + (X^T X)^{-1} X^T e$$

and

$$M_p(\hat{y}_0 | \hat{b}) = E\left[(\hat{y}_0 - X_0 b)^T(\hat{y}_0 - X_0 b)\right]$$

$$= X_0(X^T X)^{-1} X_0 \sigma^2 \tag{39}$$

(see ref. 1, p. 56).

Ridge Estimators

For the ridge estimators we recall that

$$\hat{b}(k) = (X^T X + kI)^{-1} X^T y$$

Thus,

$$\hat{y}_0 = X_0\hat{b}(k) = X_0(X^T X + kI)^{-1} X^T(Xb + e)$$

$$= X_0 Zb + X_0(X^T X + kI)^{-1} X^T e$$

where $Z = (X^TX + kI)^{-1}X^TX$. Thus,

$$M_p\left[\hat{y}_0 | \hat{b}(k)\right] = E\left[(\hat{y}_0 - X_0b)^T(\hat{y}_0 - X_0b)\right]$$

$$= E\left\{\left[X_0(X^TX + kI)^{-1}X^Te + X_0(Z - I)b\right]^T\left[X_0(X^TX + kI)^{-1}X^Te + X_0(Z - I)b\right]\right\}$$

$$= E\left[e^TX(X^TX + kI)^{-1}X_0^TX_0(X^TX + kI)^{-1}X^Te\right] + b^T(Z - I)X_0^TX_0(Z - I)b$$

$$= \gamma_1[\hat{b}(k)] + \gamma_2[\hat{b}(k)]$$

Theorem 1: The variance function $\gamma_1[\hat{b}(k)]$ is a monotonically decreasing function of k and $\gamma'[\hat{b}(0)] < 0$.

Proof: Note that if we assume $e \sim N(0, \sigma^2 I)$, then γ_1 is the expectation of a quadratic form in e. Thus (ref. 13, p. 55),

$$\gamma_1[\hat{b}(k)] = \sigma^2 \mathrm{tr}\left[X(X^TX + kI)^{-1}X_0^TX_0(X^TX + kI)^{-1}X^T\right]$$

Note that $(X^TX + kI)^{-1} = \sum_i \frac{1}{\lambda_i + k} S_iS_i^T$

and hence

$$X(X^TX + kI)^{-1}X_0^T = \sum_i \frac{1}{\lambda_i + k} XS_iS_i^TX_0^T$$

which is $n \times 1$. Thus,

$$\frac{\gamma_1[\hat{b}(k)]}{\sigma^2} = \mathrm{tr}\left[\left(\sum_i \frac{1}{\lambda_i + k} X S_i S_i^T X_0^T\right)\left(\sum_j \frac{1}{\lambda_j + k} X_0 S_j S_j^T X^T\right)\right]$$

$$= \sum_i \sum_j \frac{1}{(\lambda_i + k)(\lambda_j + k)} \mathrm{tr}\left[\left(X S_j S_j^T X_0^T\right)\left(X_0 S_i S_i^T X^T\right)\right]$$

$$= \sum_i \sum_j \frac{1}{(\lambda_i + k)(\lambda_j + k)} \left(X_0 S_i S_i^T X^T\right)\left(X S_j S_j^T X_0^T\right)$$

$$= \sum_i \sum_j \frac{1}{(\lambda_i + k)(\lambda_j + k)} X_0 \left[S_i S_i^T \left(\sum_m \lambda_m S_m S_m^T\right) S_j S_j^T\right] X_0^T$$

$$= \sum_i \frac{\lambda_i}{(\lambda_i + k)^2} X_0 S_i S_i^T X_0^T$$

Note that

$$\gamma_1[\hat{b}(0)] = \sigma^2 X_0 (X^T X)^{-1} X_0^T$$

$$\gamma_1[\hat{b}(0)] = 0$$

and

$$\gamma_1'[\hat{b}(k)] = -\sigma^2 \sum_i \frac{2\lambda_i}{(\lambda_i + k)^3} X_0 S_i S_i^T X_0^T < 0$$

Thus, γ_1 is a monotonically decreasing function of k, as was to be shown.

Theorem 2: The bias function $\gamma_2[\hat{b}(k)]$ satisfies $\gamma_2[\hat{b}(0)] = 0$ and $\gamma_2'[\hat{b}(0)] = 0$.

Proof: Recall $\gamma_2[\hat{b}(k)] = b^T(Z - I)X_0^TX_0(Z - I)b$. Since

$$Z - I = (X^TX + kI)^{-1}X^TX - I$$

$$= \sum_i \frac{-k}{\lambda_i + k} S_iS_i^T$$

we obtain

$$\gamma_2[\hat{b}(k)] = \sum_i \sum_j \frac{k^2}{(\lambda_i + k)(\lambda_j + k)} b^TS_iS_i^TX_0^TX_0S_jS_j^Tb$$

Let

$$f_{ij}(k) = \frac{k^2}{(\lambda_i + k)(\lambda_j + k)}$$

Then

$$f'_{ij}(k) = \frac{k^2(\lambda_i + \lambda_j) + 2k\lambda_i\lambda_j}{(\lambda_i + k)^2(\lambda_j + k)^2} \geq 0$$

Thus, it is easily seen that $\gamma_2[\hat{b}(0)] = 0$ and $\gamma'_2[\hat{b}(0)] = 0$.

Theorem 3: $M_p\left[\hat{y}_0 \middle| \hat{b}(k)\right]$ is initially decreasing in k.

Proof: The result follows directly from theorems 1 and 2.

From theorem 3 we thus have the result that it is theoretically possible to reduce the mean squared error by slightly increasing k. We have the same drawbacks as for ridge estimators. Namely, how to specify better estimators is unknown; and they will typically be functions of σ^2 and b, which are assumed to be unknown.

Marquardt's Generalized Inverse Estimators

For these estimators we recall that

$$\hat{b}(\rho) = G_\rho X^T y = (X^T X)^+_\rho X^T y$$

and

$$H_\rho = G_\rho X^T X = \sum_{j=1}^{\rho*} s_j s_j^T + d\rho \; s_{\rho*+1} s_{\rho*+1}^T$$

Then since

$$y = Xb + e$$

$$\hat{y}_0 = X_0 \hat{b}(\rho) = X_0 G_\rho X^T y = X_0 G_\rho X^T (Xb + e) = X_0 H_\rho b + X_0 G_\rho X^T e \qquad (40)$$

The expected mean squared error of the predicted regression function is

$$M_p\left[\hat{y}_0 \middle| \hat{b}(\rho)\right] = E\left[(\hat{y}_0 - X_0 b)^T (\hat{y}_0 - X_0 b)\right]$$

$$= E\left\{\left[X_0 G_\rho X_0^T e + X_0 (H_\rho - I) b\right]^T \left[X_0 G_\rho X^T e + X_0 (H_\rho - I) b\right]\right\}$$

$$= E\left(e^T X G_\rho X^T X_0 G_\rho X^T e\right) + b^T (H_\rho - I) X_0^T X_0 (H_\rho - I) b$$

$$= \gamma_1[\hat{b}(\rho)] + \gamma_2[\hat{b}(\rho)]$$

Theorem 4: The variance function $\gamma_1[\hat{b}(\rho)]$ is a monotonically increasing function of ρ and $\gamma_1'[\hat{b}(\rho)] > 0$.

Proof: Since we assume $e \sim N(0, \sigma^2 I)$, we have that

$$\gamma_1[\hat{b}(\rho)] = \sigma^2 \mathrm{tr}\left(X G_\rho X_0^T X_0 G_\rho X^T\right)$$

From equation (26) for G_ρ we may show

$$X G_\rho X_0^T = \sum_{i=1}^{\rho*} \frac{1}{\lambda_i} X s_i s_i^T X_0^T + \frac{(d\rho)}{\lambda_{\rho*+1}} X s_{\rho*+1} s_{\rho*+1}^T X_0^T$$

and $XG_\rho X_0^T$ is an $n \times 1$ vector. Thus,

$$\mathrm{tr}\left[\left(XG_\rho X_0^T\right)\left(X_0 G_\rho X^T\right)\right] = \left(X_0 G_\rho X^T\right)\left(XG_\rho X_0^T\right) = \sum_i \sum_j \frac{1}{\lambda_i \lambda_j} X_0 S_i S_i^T (X^T X) S_j S_j^T X_0^T$$

$$+ \sum_i \frac{1}{\lambda_i} X_0 S_i S_i^T X^T X S_{\rho*+1} S_{\rho*+1}^T X_0 + \sum_j \frac{1}{\lambda_j} X_0 S_{\rho*+1} S_{\rho*+1}^T X^T X S_j S_j^T X_0$$

$$+ \frac{(d\rho)^2}{\lambda_{\rho*+1}} X_0 S_{\rho*+1} S_{\rho*+1}^T X^T X S_{\rho*+1} S_{\rho*+1}^T X_0$$

Hence,

$$\gamma_1[\hat{b}(\rho)] = \sum_{i=1}^{\rho*} \frac{\sigma^2}{\lambda_i} X_0 S_i S_i^T X_0^T + \frac{(\sigma d\rho)^2}{\lambda_{\rho*+1}} X_0 S_{\rho*+1} S_{\rho*+1}^T X_0^T$$

and

$$\gamma_1[\hat{b}(0)] = 0$$

This function is continuous from $\rho = 0$ to $\rho = p$, and the derivative is continuous for nonintegral ρ. Between the two integers $\rho*$ and $\rho* + 1$ we have

$$\gamma_1'[\hat{b}(\rho)] = \frac{2\sigma^2 d\rho}{\lambda_{\rho*+1}} X_0 S_{\rho*+1} S_{\rho*+1}^T X_0^T > 0$$

Theorem 5: The bias function $\gamma_2[\hat{b}(\rho)]$ satisfies $\gamma_2[\hat{b}(p)] = 0$ and $\gamma_2'[\hat{b}(p)] = 0$.
Proof: We have

$$\gamma_2[\hat{b}(\rho)] = b^T (H_\rho - I) X_0^T X_0 (H_\rho - I) b \qquad (41)$$

where

$$H_\rho - I = -\sum_{j=\rho^*+2}^{p} s_j s_j^T + (d\rho - 1)s_{\rho^*+1}s_{\rho^*+1}^T$$

and the sum is null if $\rho^* + 2 > p$. Thus, equation (41) may be expressed as

$$\gamma_2[\hat{b}(\rho)] = \sum_{i=\rho^*+2}^{p}\ \sum_{j=\rho^*+2}^{p} b^T s_i s_i^T X_0^T X_0 s_j s_j^T b - \sum_{i=\rho^*+2}^{p} (d\rho - 1)b^T s_{\rho^*+1}s_{\rho^*+1}^T X_0^T X_0 s_i s_i^T b$$

$$- \sum_{j=\rho^*+2}^{p} (d\rho - 1)b^T s_j s_j^T X_0^T X_0 s_{\rho^*+1}s_{\rho^*+1}^T b$$

$$+ (d\rho - 1)^2 b^T s_{\rho^*+1}s_{\rho^*+1}^T X_0^T X_0 s_{\rho^*+1}s_{\rho^*+1}^T b$$

From the preceding, it may be verified that, since $d\rho$ approaches 1 linearly with ρ, $\gamma_2[\hat{b}(p)] = 0$. Upon differentiating with respect to $d\rho$, we also may verify that $\gamma_2'[\hat{b}(p)] = 0$.

Theorem 6: The mean square error function $M_p\left[\hat{y}_0 \middle| \hat{b}(\rho)\right]$ is initially decreasing as ρ decreases from $\rho = p$.

Proof: Immediate from theorems 4 and 5.

Theorem 6 indicates that it is possible to improve the mean squared error of the predicted regression, at least initially, by decreasing ρ. However, we are unable to specify how much to decrease ρ. Also, optimal values of ρ (if they exist) would be expected to be functions of σ^2 and b, which are assumed to be unknown.

Shrunken Estimators

For any of the shrunken estimators, we have that

$$\hat{b}(c) = c\hat{b} \qquad 0 \leq c \leq 1$$

and hence

$$\hat{y}_0 = cX_0\hat{b} = cX_0(X^TX)^{-1}X^T(Xb + e)$$

Thus,

$$
\begin{aligned}
M_p\left[\hat{y}_0 \middle| \hat{b}(c)\right] &= E\left[(\hat{y}_0 - X_0 b)^T(\hat{y}_0 - X_0 b)\right] \\
&= E\left\{\left[(c - 1)X_0 b + cX_0(X^TX)^{-1}X^Te\right]^T\left[(c - 1)X_0 b + cX_0(X^TX)^{-1}X^Te\right]\right\} \\
&= E\left[c^2e^TX(X^TX)^{-1}X_0^TX_0(X^TX)^{-1}X^Te\right] + 2E\left[(c - 1)cb^TX_0^TX_0(X^TX)^{-1}X^Te\right] \\
&\qquad + E\left[(c - 1)^2b^TX_0^TX_0b\right]
\end{aligned}
$$

For stochastically shrunken estimators, these expectations may be somewhat difficult. For a deterministically shrunken estimator, c is a constant and $M_p[\hat{b}(c)]$ is easily found to be

$$
\begin{aligned}
M_p\left[\hat{y}_0 \middle| \hat{b}(c)\right] &= c^2E\left[e^TX(X^TX)^{-1}X_0^TX_0(X^TX)^{-1}X^Te\right] + (c - 1)^2b^TX_0^TX_0b \\
&= \gamma_1[\hat{b}(c)] + \gamma_2[\hat{b}(c)]
\end{aligned}
$$

Theorem 7: The variance function $\gamma_1[\hat{b}(c)]$ is a monotonically increasing function of $c > 0$ and $\gamma_1'[\hat{b}(1)] \geq 0$.

Proof: Since γ_1 is simply a positive constant times c^2, it is immediately seen that γ_1 satisfies the stated conditions.

Theorem 8: The bias function $\gamma_2[\hat{b}(c)]$ is a monotonically decreasing function of c for $0 \leq c \leq 1$.

Proof: Since γ_2 is simply a positive constant times $(c - 1)^2$, the result follows immediately.

Theorem 9: $M_p\left[\hat{y}_0 \middle| \hat{b}(c)\right]$ is initially decreasing as c decreases from $c = 1$, and there is a unique minimum for some $0 < c < 1$.

Proof: Immediate from theorems 7 and 8.

Theorem 9 states that an optimal choice of c exists. However, this optimal value of c will be a function of σ^2 and b. It should be noted that the stochastically shrunken estimator given by equation (31) (i.e., the estimator discussed by Sclove) seems to be the most likely one for which some optimality property of the predicted regression function will be achieved.

Principal Components Estimators

The principal components estimators of the second type depend upon the particular method used for determining the significance of the coefficients. We will not consider the mean squared error of the predicted regression function in this report. Kennedy and Bancroft (ref. 19) and Holms (ref. 20) have considered two different procedures for subset regression in the principal components case. Both references are concerned with the prediction problem. There is no clear way of comparing their results to our results since most of their results are based on Monte-Carlo simulation studies.

Discussion

From the preceding developments we can draw the conclusion that all the biased estimators discussed offer the possibility of decreased mean squared error of the estimated regression function. This possibility can be realized only in the event that we can identify particular better members of each class. At the current state of the art, there is no way known to do this. The two most promising possibilities seem to be principal components and Sclove's shrunken estimator. Principal components is appealing because of some Monte-Carlo simulation work reported in references 19 and 20. Sclove's estimator is promising in the sense that the estimator for the regression coefficients can be proven to have smaller mean squared error than the least-squares estimator.

HYPOTHESIS TESTING

The third major objective of a linear model analysis is to determine if some of the parameters in the model can reasonably be set to zero. To do this, the distributional properties of the estimators must be known in order to perform significance tests. For the ridge, generalized inverse, and shrunken estimators, this information is not available. Thus, hypothesis testing is not possible. For the principal components estimators, distributional properties of the estimators of the transformed model are better known. However, an experimenter is more interested in hypothesis testing in terms of the original parameterization. And it is most unlikely that a subset regression in the transformed model will also correspond to a subset regression in the original parameterization.

Ordinary least squares seems to be the only method available for subset regression in the original parameterization. Simultaneous deletion of variables (and parameters) based on confidence ellipsoids (for parameters) would be a procedure with a good statistical basis. The more usual subset regression procedures based on stepwise regression, or some such, have the drawback that they involve nonindependent repeated significance tests. Theoretical developments are difficult, but there has been much work in the area based on simulations. To this author, the usual least-squares subset regression procedures are the most appealing.

EXAMPLES

Example 1

We use the example studied in some detail by Marquardt (ref. 7). The linear regression model is

$$E(y) = b_1x_1 + b_2x_2$$

and the estimated model is

$$\hat{y} = \hat{b}_1^* x_1 + \hat{b}_2^* x_2$$

where the $\hat{b}_i^*$ are to be estimated by several methods for comparison. The observed data are

$$X = \begin{bmatrix} 3/5\sqrt{1/2} & 4/5\sqrt{1/2} \\ 4/5\sqrt{1/2} & 3/5\sqrt{1/2} \\ 5/5\sqrt{1/2} & 5/5\sqrt{1/2} \end{bmatrix}$$

$$y = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

The eigenvalues and associated normalized eigenvectors of X^TX and $(X^TX)^{-1}$ are

$$X^TX : \begin{cases} \lambda_1 = 1.98 & S_1^T = \left(\sqrt{2}/2,\ \sqrt{2}/2\right) \\ \lambda_2 = 0.02 & S_2^T = \left(\sqrt{2}/2, -\sqrt{2}/2\right) \end{cases}$$

$$(X^TX)^{-1} : \begin{cases} \lambda_1 = 0.505 & S_1^T = \left(\sqrt{2}/2,\ \sqrt{2}/2\right) \\ \lambda_2 = 50.00 & S_2^T = \left(\sqrt{2}/2, -\sqrt{2}/2\right) \end{cases}$$

It is clear from these eigenspace descriptions that X^TX has the formal rank of 2 but is essentially of rank 1. In the parameter space, the variance of the linear function $\sqrt{2}/2(b_1 + b_2)$ is about 1 percent of the total variance, whereas $\sqrt{2}/2(b_1 - b_2)$ has about 99 percent of the total variance. This is equivalent to saying that $\sqrt{2}/2(b_1 + b_2)$ is well determined, while $\sqrt{2}/2(b_1 - b_2)$ is not. We drop the factors $\sqrt{2}/2$ and thus consider only $b_1 + b_2$ and $b_1 - b_2$ for simplicity.

Marquardt performed ridge regressions for various values of k, generalized inverse regressions for various values of ρ, and the two best subset regressions. Some of the results of these analyses are presented in table III.

Table III(a) presents the ridge regression results for several values of k given in the first column. The next two columns give the individual estimates of $\hat{b}_1$ and $\hat{b}_2$. The following columns give the estimates for $\hat{b}_1 + \hat{b}_2$ and $\hat{b}_1 - \hat{b}_2$. The values of k, $\hat{b}_1$, and $\hat{b}_2$ are an abbreviated set of results taken directly from Marquardt. The estimates $\hat{b}_1 + \hat{b}_2$ and $\hat{b}_1 - \hat{b}_2$ are not given by Marquardt. It may be noted that $\hat{b}_1 - \hat{b}_2$ is much more rapidly decreasing as k increases than is $\hat{b}_1 + \hat{b}_2$. This is in accord with theory (eq. (22)).

Table III(b) is more interesting inasmuch as generalized inverse regression has a more direct relation to estimable function concepts than does ridge regression. The format of the table is identical to that of table III(a). It is interesting to note that for $1 \leq \rho \leq 2$, $\hat{b}_1 + \hat{b}_2$ is unchanged while $\hat{b}_1 - \hat{b}_2$ decreases linearly in ρ to the value zero for $\rho = 1.0$. Then for $0 \leq \rho \leq 1$, $\hat{b}_1 - \hat{b}_2$ remains at zero, while $\hat{b}_1 + \hat{b}_2$ decreases linearly to zero as ρ decreases. This is in accord with theory (eq. (27)).

Table III(c) presents the results for the two subset regressions using x_1 only or x_2 only. The comparison of $\hat{b}_1 + \hat{b}_2$ and $\hat{b}_1 - \hat{b}_2$ resulting from these two regressions to the corresponding values from the full regression is most instructive. Note that the eigenspace analysis indicates $\hat{b}_1 + \hat{b}_2$ should be well estimated and has an estimated value of 3.6427 from the full regression. The corresponding values from the subset regressions are both quite close. But the two values for $\hat{b}_1 - \hat{b}_2$ are not close and in fact the x_1-only regression yields a value much closer to the full regression value. Based upon this and the comparison between the two residual sums of squares the x_1-only subset regression is clearly superior.

It should also be noted that Marquardt chooses the ridge estimator with $k = 0.2$ as the best ridge estimator. This yields a residual sum of squares of 0.887. He also chooses the generalized inverse estimator with $\rho = 1$ as the best of that class of estimators. This yields a residual sum of squares of 0.864. The x_1-only subset regression, however, provides a residual sum of squares of 0.480. The x_1-only subset regression also comes within ϵ of not violating the stipulation that both regression parameters are expected, because of physical considerations, to be positive.

We would consider the x_1-only subset regression to be superior to either the ridge or generalized inverse estimators since it fits the data better and has a more readily interpretable constraint.

Example 2

For our second example we consider the data of Gorman and Toman (ref. 21), which were used as an illustrative example by Hoerl and Kennard (ref. 5). The data are presented in table IV in correlation form. Table V presents the eigenvalues and eigenvectors of X^TX. Although we used the two-digit correlations in all calculations just as Hoerl and Kennard did, slightly different eigenvalues were obtained. Our calculations were done by using the double-precision version of the IBM SSP EIGEN subroutine (ref. 22) on an IBM 360/67.

Using the data matrix of table IV, we performed ridge regressions for several values of k and generalized inverse regressions for several integral values of ρ. These results are presented in tables VI and VII, respectively.

Table VI provides the following information about the ridge regression results: The first column indicates the various values of k for which ridge regressions were calculated. For these values of k, the coefficient vectors $\hat{b}(k)$ were calculated. These are not provided in the table. Instead, the values of $S_i^T\hat{b}(k)$ were calculated. These values are provided in the next 10 columns, for $i = 1$ to 10. The last column gives the residual sum of squares for each value of k. Upon examination of the table, it may readily be seen that $S_i^T\hat{b}_i(k)$ decreases as k increases. The rate of decrease is most rapid for $i = 10$ and least rapid for $i = 1$. This is in accord with the theory previously discussed (eq. (22)). The bottom row of the table presents the estimator $\hat{b}(k = 0.25)$ which Hoerl and Kennard chose as their "optimum" ridge estimator.

Table VII provides the following information about the generalized inverse regressions: The first column provides the values of ρ for which generalized inverse regressions were performed. For these values of ρ, the coefficients $\hat{b}_i(\rho)$ were calculated, and these are presented in the next 10 columns, for $i = 1$ to 10. The last column provides the residual sum of squares for each value of ρ. The bottom row of the table provides the values of $S_i^T\hat{b}(\rho = 10)$. As indicated previously, these represent what might be called the estimable functions. The effect of changing ρ is to leave $S_i^T\hat{b}(\rho)$ unchanged for $i \leq \rho$ and to set $S_i^T\hat{b}(\rho) = 0$ for $i > \rho$. We thus need only to present the $S_i^T\hat{b}(\rho = 10)$ values for comparison purposes.

Generalized inverse estimators have not previously been applied to these data in the literature. For comparison purposes we will simply choose $\hat{b}(\rho = 9)$ as the "optimal" generalized inverse estimator since it provides very nearly the same residual sum of squares as the "optimal" ridge estimator. This choice also provides a maximum variance inflation factor of 3.85.

It may be noted that there is no consistent behavior of the individual components of $\hat{b}(\rho)$ as ρ decreases from $\rho = 10$ to $\rho = 6$. That there is a consistency in $\hat{b}(\rho)$ as a whole, though, is evidenced by the invariance of the $S_i^T\hat{b}(\rho)$ values.

When Gorman and Toman studied these data, they arrived at a best subset regression in which variables 1, 4, 9, and 10 were deleted. We will not consider here that there might be a better subset regression (as indeed there might be). The individual parameter estimates for the chosen subset are presented in table VIII. Also given is the residual sum of squares for the subset regression. For purposes of comparison, let $\hat{b}(d)$ denote the estimate of b where we have deleted variables 1, 4, 9, and 10. In effect we have set $b_1 = b_4 = b_9 = b_{10} = 0$. The table also provides the values of $S_i^T\hat{b}(d)$ and $S_i^T\hat{b}$ for comparison purposes.

We consider three criteria to compare the three "optimal" estimators. The first criterion is the residual sum of squares value. The second criterion is defined as follows: Let

$$L^T(\hat{b}*) = \left(S_1^T\hat{b}*, \ldots, S_p^T\hat{b}*\right)$$

where $\hat{b}*$ is some estimator of b. Now compute the variance-weighted squared distance of each of $L[\hat{b}(k)]$, $L[\hat{b}(\rho)]$, and $L[\hat{b}(d)]$ from $L(\hat{b})$. That is, let

$$d_k^2 = \sum_{i=1}^{p} \lambda_i\left[S_i^T\hat{b}(k) - S_i^T\hat{b}\right]^2$$

$$d_\rho^2 = \sum_{i=1}^{p} \lambda_i\left[S_i^T\hat{b}(\rho) - S_i^T\hat{b}\right]^2$$

and

$$d_d^2 = \sum_{i=1}^{p} \lambda_i \left[S_i^T \hat{b}(d) - S_i^T \hat{b} \right]^2$$

The third criterion is difficult to quantify. It is the number and type of constraint placed upon the parameter space to obtain the estimator. For instance, $\hat{b}$ places no constraints upon the parameter space. The estimator $\hat{b}(d)$ places certain restraints upon the parameter space by restricting certain components of b to zero. The generalized inverse estimators impose that certain linear combinations of $\hat{b}$ be set to zero. They also impose the restraint that $\hat{b}(\rho)$ be a minimum-norm solution to the modified normal equations. The ridge estimators impose the constraint that $\hat{b}(k)$ be the minimum-norm estimator in a class of estimators defined by hyperellipsoids in the parameter space.

Table IX presents the three criteria descriptions for the "optimal" members of the ridge, generalized inverse, and subset regression estimators. The comparison of these criteria cannot be entirely objective. On the basis of residual sums of squares and the constraints, it would seem that the subset regression is best. The residual sum of squares is lowest, and there are four constraints which are quite easy to interpret. The subset regression, however, has a larger d. This indicates that dropping variables 1, 4, 9, and 10 has affected the estimable functions more. One interpretation of this larger distance is that the estimator $\hat{b}(d)$ is farther from the center of the hyperellipsoids which define confidence regions for b based on the full least-squares solution. A closer subset regression is that in which only variable 4 is dropped. This yields a residual sum of squares of 0.115 and a distance d of 1.04.

Considering the fact that the subset regression with variables 1, 4, 9, and 10 dropped involves four constraints as opposed to the one constraint on $\hat{b}(\rho = 9)$ or the difficult-to-characterize constraint of $\hat{b}(k = 0.25)$, the performance of the usual subset regression procedures seems quite good.

CONCLUSIONS

We have considered the five major types of biased estimators that have been proposed in the literature. These are ridge, Marquardt's generalized inverse, shrunken, principal components, and subset regression.

We present the biased and unbiased estimators of the parameters in a linear model. The presentation centers on a duality of the X^TX matrix of the least-squares normal equations. The duality is in the sense that X^TX in its eigenspace representation describes how and how well the data space is covered, while the similar representation of $(X^TX)^{-1}$ describes how the distribution of the estimated parameters is spread out in the parameter space.

We consider biased estimators with respect to all three major objectives of a linear model analysis; that is, point estimation of the parameters, estimation of the predictive regression function, and hypothesis testing of the parameters. Our major conclusions with respect to these objectives are as follows:

Estimation of Parameters

1. In a nearly singular system, the full parameter vector is essentially inestimable. However, certain linear combinations of the parameters are estimable.

2. Biased estimators all place some kind of constraints on the parameter space in order to achieve "better" estimators.

3. The decision to use mean squared error as a criterion of goodness should be made independently of the condition of X^TX.

4. If mean squared error is to be accepted as a criterion of goodness, then only one estimator so far proposed (i.e., Sclove's) has any proven optimality properties.

5. Because of the lack of distributional information, no biased estimator provides interval estimation capability.

Estimation of Regression Function

All the biased estimators discussed offer the possibility of decreased mean squared error of the predicted regression function. This possibility cannot be assured (except for two special cases of principal components estimators) because it is not known how to identify the members of each class of biased estimators that provide smaller mean squared error.

Hypothesis Testing

Only the ordinary least-squares estimators have enough of the distributional theory available to provide subset regression techniques in the original parameterization.

The overall conclusion is that ordinary least-squares estimation and subset regression methods are still the preferred methods of linear model analysis in the regression situation.

Lewis Research Center,
National Aeronautics and Space Administration,
Cleveland, Ohio, January 14, 1975,
506-21.

REFERENCES

1. Draper, N. R.; and Smith, H.: Applied Regression Analysis. John Wiley & Sons, Inc., 1966.

2. Stein, C.: Multiple Regression. Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, I. Olkin, ed., Stanford Univ. Press, 1960, pp. 424-443.

3. James, W.; and Stein, C.: Estimation with Quadratic Loss. Proc. of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. I, Univ. Cal. Press, 1961, pp. 361-379.

4. Sclove, Stanley L.: Improved Estimators for Coefficients in Linear Regression. J. American Statist. Assoc., vol. 63, 1968, pp. 596-606.

5. Hoerl, Arthur E.; and Kennard, Robert W.: Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, vol. 12, no. 1, Feb. 1970, pp. 55-67.

6. Hoerl, Arthur E.; and Kennard, Robert W.: Ridge Regression: Applications to Nonorthogonal Problems. Technometrics, vol. 12, no. 1, Feb. 1970, pp. 69-82.

7. Marquardt, Donald W.: Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation. Technometrics, vol. 12, no. 3, Aug. 1970, pp. 591-612.

8. Mayer, Lawrence S.; and Willke, Thomas A.: On Biased Estimation in Linear Models. Technometrics, vol. 15, no. 3, Aug. 1973, pp. 497-508.

9. Kendall, M. G.: A Course in Multivariate Analysis. Griffin and Co., Ltd. (London), 1957.

10. Massy, William F.: Principal Components Regression in Exploratory Statistical Research. J. American Statist. Assoc., vol. 60, 1965, pp. 234-256.

11. Newhouse, Joseph P.; and Oman, Samuel D.: An Evaluation of Ridge Estimators. R-716-PR, Rand Corp. (AD-723626), 1971.

12. Searle, S. R.: Matrix Algebra for the Biological Sciences. John Wiley & Sons, Inc., 1966.

13. Searle, S. R.: Linear Models. John Wiley & Sons, Inc., 1971.

14. Federer, W. T.: Experimental Design. MacMillan Co., 1955.

15. Sidik, Steven M.: An Improved Multiple Linear Regression and Data Analysis Computer Program Package. NASA TN D-6770, 1972.

16. Rao, Colyompudi R.: Linear Statistical Inference and Its Applications. John Wiley & Sons, Inc., 1965.

17. Kendall, M. G.; and Stuart, A.: The Advanced Theory of Statistics. Vol. 3, Hafner Publ. Co., 1966.

18. Rao, Colyompudi R.; and Mitra, S. K.: Generalized Inverse of Matrices and Its Applications. John Wiley & Sons, Inc., 1971.

19. Kennedy, W. J.; and Bancroft, T. A.: Model Building for Prediction in Regression Based upon Repeated Significance Tests. Annals of Math. Statist., vol. 42, no. 4, Aug. 1971, pp. 1273-1284.

20. Holms, Arthur G.: "Chain Pooling" to Minimize Prediction Errors in Subset Regression. NASA TM X-71645, 1974.

21. Gorman, J. W.; and Toman, R. J.: Selection of Variables for Fitting Equations to Data. Technometrics, vol. 8, no. 1, Feb. 1966, pp. 27-51.

22. Systems/360 Scientific Subroutine Package (360A-CM-03X) Version II Programmer's Manual. Technical Publications Dept., IBM Corp., 1967.

TABLE I. - WEIGHTS OF RUBBER PLANTS

Normal	Off-type	Aberrant
$y_{11} = 101$ $y_{12} = 105$ $y_{13} = 94$	$y_{21} = 84$ $y_{22} = 88$	$y_{31} = 32$

TABLE II. - THREE INDEPENDENT ESTIMABLE FUNCTIONS

w_1	w_2	w_3	Function	Estimator
1	-1	0	$\alpha_1 - \alpha_2$	$\bar{y}_{1.} - \bar{y}_{2.} = 14$
0	1	-1	$\alpha_2 - \alpha_3$	$\bar{y}_{2.} - \bar{y}_{3.} = 54$
1/3	1/3	1/3	$\mu + 1/3(\alpha_1 + \alpha_2 + \alpha_3)$	$1/3(\bar{y}_{1.} + \bar{y}_{2.} + \bar{y}_{3.}) = 72\frac{2}{3}$

TABLE III. - RESULTS OF REGRESSIONS

FOR EXAMPLE 1

(a) Ridge regression

k	$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_1 + \hat{b}_2$	$\hat{b}_1 - \hat{b}_2$	Residual sum of squares
0.00	5.3569	-1.7142	3.6427	7.0711	0.3636
.02	3.5709	.0354	3.6063	3.5355	.490
.04	2.9638	.6068	3.5706	2.3570	.591
.10	2.3230	1.1445	3.4675	1.1785	.741
.20	1.9757	1.3328	3.3085	.6429	.887
.40	1.6836	1.3469	3.0305	.3367	1.188
1.00	1.2795	1.1408	2.4203	.1387	2.324

(b) Generalized inverse regression

ρ	$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_1 + \hat{b}_2$	$\hat{b}_1 - \hat{b}_2$	Residual sum of squares
2.0	5.3569	-1.7142	3.6427	7.0711	0.3636
1.6	3.9427	-.3000	3.6427	4.2427	.444
1.2	2.5284	1.1142	3.6426	1.4142	.684
1.0	1.8213	1.8213	3.6426	0	.864
.5	.9107	.9107	1.8214	0	4.148

(c) Best subset regression

Subset	$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_1 + \hat{b}_2$	$\hat{b}_1 - \hat{b}_2$	Residual sum of squares
x_1	3.6770	0	3.6770	3.6770	0.4800
x_2	0	3.5355	3.5355	-3.5355	1.5000

TABLE IV. - DATA MATRIX FOR EXAMPLE 2 IN CORRELATION FORM

[All calculations used two-digit correlations]

X^TX	1	2	3	4	5	6	7	8	9	10
1	1.0									
2	-.04	1.0								
3	.51	.0	1.0							
4	.12	-.16	0	1.0						
5	-.71	.06	-.59	-.07	1.0					
6	-.87	.09	-.65	-.09	.84	1.0				
7	-.09	.24	-.02	.03	.38	.13	1.0			
8	0	.01	.34	.08	-.36	-.20	-.48	1.0		
9	-.09	.09	-.08	.02	-.14	.04	.07	-.18	1.0	
10	-.36	-.30	-.44	-.09	.54	.45	.40	-.46	.05	1.0
X^Ty	-0.81	-0.10	-0.63	-0.10	0.56	0.81	0.04	0.06	0.16	0.45

TABLE V. - SPECTRAL DECOMPOSITION OF X^TX

Eigenvalues, λ_i	Components of associated eigenvectors, S_i, for i =									
	1	2	3	4	5	6	7	8	9	10
3.694	0.408	-0.014	0.387	0.064	-0.473	-0.465	-0.206	0.250	-0.036	-0.365
1.533	.364	.050	.132	.034	-.068	-.275	.574	-.570	.217	.254
1.294	-.108	.806	.093	-.361	.027	.087	.191	.063	.177	-.345
1.054	-.106	-.026	-.209	.302	-.205	.059	-.216	-.026	.868	-.079
.971	-.027	.242	.019	.854	.155	.063	.292	.141	-.197	-.199
.668	-.349	-.228	.612	-.057	.022	.051	.372	.440	.246	.231
.358	.178	.291	-.414	.001	-.228	-.191	.100	.518	-.017	.588
.220	-.078	-.368	-.482	-.187	-.013	-.208	.517	.216	.031	-.482
.137	.591	-.063	.047	-.067	.701	.038	-.078	.272	.252	-.050
.070	.410	-.107	.009	-.044	-.400	.0781	.193	.074	-.039	-.065

TABLE VI. - RESULTS OF RIDGE REGRESSIONS FOR EXAMPLE 2

k	Estimable functions, $S_i^T\hat{b}(k)$, for i =										Residual sum of squares
	1	2	3	4	5	6	7	8	9	10	
0	-0.375	-0.332	-0.035	0.206	-0.065	0.245	0.287	0.311	-0.300	0.877	0.105
.02	-.373	-.328	-.034	.202	-.064	.238	.271	.285	-.262	.683	.108
.04	-.371	-.324	-.034	.198	-.062	.231	.258	.263	-.232	.559	.114
.06	-.369	-.320	-.033	.195	-.061	.224	.245	.244	-.209	.473	.120
.08	-.367	-.316	-.033	.191	-.060	.218	.234	.228	-.190	.410	.126
.10	-.365	-.312	-.032	.188	-.059	.213	.224	.213	-.174	.362	.131
.15	-.360	-.302	-.031	.180	-.056	.200	.202	.185	-.143	.280	.144
.20	-.356	-.294	-.030	.173	-.054	.188	.184	.163	-.122	.228	.154
.25*	-.351	-.286	-.029	.166	-.052	.178	.169	.145	-.106	.192	.164
.30	-.347	-.278	-.028	.160	-.050	.169	.156	.131	-.094	.166	.173
.50	-.330	-.250	-.025	.139	-.043	.140	.120	.095	-.065	.108	.204
1.00	-.295	-.201	-.020	.106	-.032	.098	.076	.056	-.036	.056	.268
$\hat{b}(0.25)^*$	-0.288	-0.109	-0.246	-0.054	-0.045	0.339	0.055	0.244	0.111	0.126	0.164

TABLE VII. - RESULTS OF GENERALIZED INVERSE REGRESSIONS FOR EXAMPLE 2

ρ	Regression coefficients, $\hat{b}_i(\rho)$, for i =										Residual sum of squares
	1	2	3	4	5	6	7	8	9	10	
10	-0.166	-0.223	-0.361	-0.106	-0.479	0.837	0.289	0.385	0.082	0.094	0.105
9	-.526	-.128	-.369	-.068	-.127	.152	.120	.320	.116	.151	.159
8	-.349	-.147	-.355	-.088	.083	.164	.096	.402	.192	.136	.171
7	-.325	-.033	-.205	-.030	.087	.228	-.064	.335	.182	.286	.192
6	-.376	-.116	-.086	-.030	.152	.283	-.093	.186	.187	.118	.222
$S_i^T\hat{b}(p)$	-0.375	-0.332	-0.035	0.206	-0.065	0.245	0.287	0.311	-0.300	0.877	

TABLE VIII. - SUBSET REGRESSION

WITH VARIABLES 1, 4, 9,

AND 10 DELETED

[Residual sum of squares, 0.137.]

i	Parameter estimates, $\hat{b}_i$	Estimable functions	
		$S_i^T\hat{b}(d)$	$S_i^T\hat{b}$
1	------	-0.382	-0.375
2	-0.249	.725	-.332
3	-.367	.267	-.035
4	------	-.472	.206
5	-.546	-.135	-.065
6	1.08	-.084	.245
7	.336	.254	.287
8	.369	.771	.311
9	------	.260	-.300
10	------	.441	.877

TABLE IX. - OPTIMAL ESTIMATORS AND CRITERIA OF COMPARISON

Type	Residual sum of squares	Distance, d	Constraints
Subset (drop variables 1, 4, 9, and 10)	0.137	1.63	$b_1 = b_4 = b_9 = b_{10} = 0$
Generalized inverse, $\hat{b}(\rho = 9)$	.159	.48	Minimum-norm estimator such that $S_{10}^T b = 0$
Ridge, $\hat{b}(k = 0.25)$	.164	.49	------------------------
Subset (drop variable 4)	.115	1.04	$b_4 = 0$

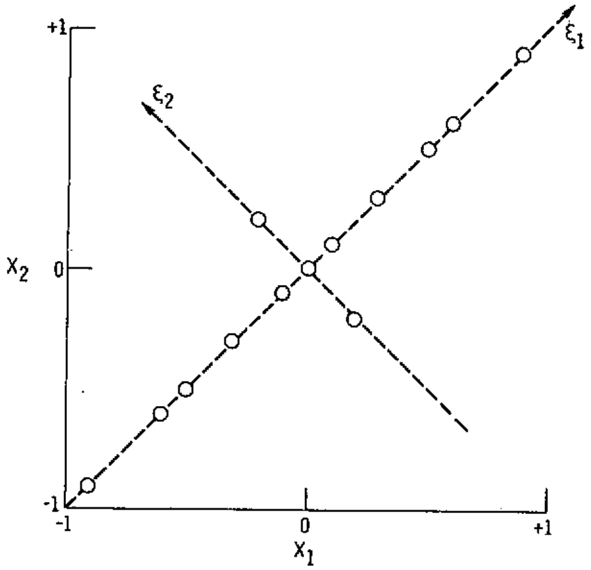

Figure 1. - Two-dimensional example for duality of X^TX.